CREDIT EDA Assignment
DS C40 Batch
Module 7

Presented By:
Avin Seth

Contents

- Problem Statement/Business Objective
- Road Taken
- Data Analysis
  - Univariate
  - Bivariate
  - Multivariate

Business Understanding

- When the company receives a loan application, the company has to decide for loan approval based on the applicant’s profile. Two types of risks are associated with the bank’s decision:
  - If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
  - If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.
- There are 4 scenarios that happen when a loan is applied:
  - Loan is Approved
  - Loan is Cancelled
  - Loan is Refused
  - Unused Offer

Problem Statement/Business Objective

- **Primary Aim:** Identifying various factors in BFSI that lead to default payments in loans, so that at the time of lending money to customers, the bank or the financial institution does not face any money loss. The company can utilise this knowledge for its portfolio and risk assessment.

- **Identifying Patterns:** Indicates if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

The Road Ahead

Steps Taken

- **Primary data inspection**
  - 1st step taken was to import the libraries in the notebook
  - 2nd step is to load the csv files
  - 3rd step check data info
  - 4th step check the Mathematical parameters of the dataset
- **Step 1:Identify the Missing Values and Rectify them**
  - Dropped columns with missing values>45% in both the datasets
  - Dropped other columns that does not put impact on the target column
  - Imputing Missing values
  - Correcting Data types if any
  - Rectifying the negative values

The Road Ahead

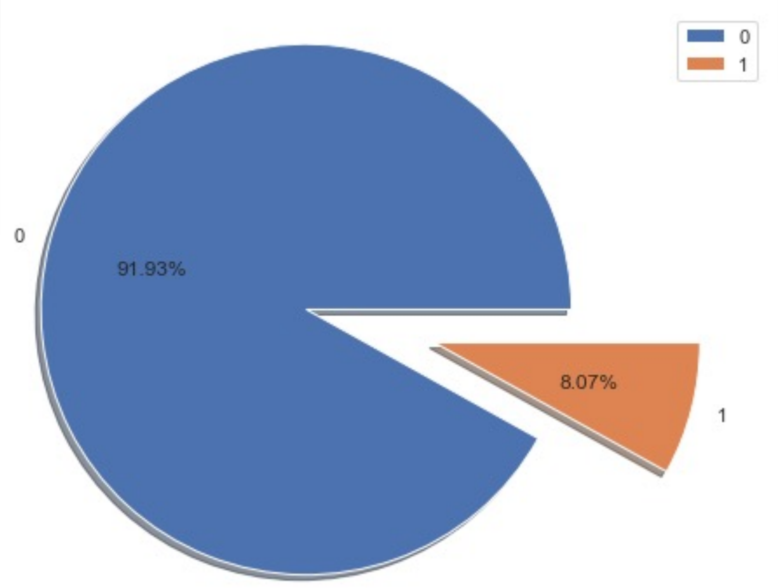
Steps Taken Contd..

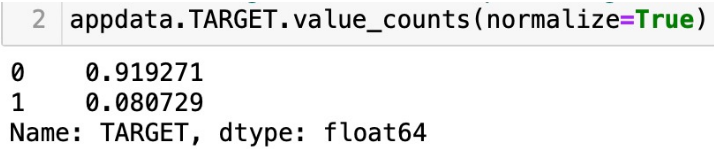
- Step 2: Check for Outliers and do the necessary treatment of outliers
- Step 3: Binning of certain columns into categories
- Step 4: Calculate imbalance percentage for target column
- Step 5: Analysis
  - Univariate Analysis
  - Bivariate Analysis
  - Multivariate Analysis

Analysis

Imbalance Percentage

This Pie plot on target variable indicates that out of the total data the number of defaulters are very less (8.07% of the total data) as compared the non defaulters (91.93% of the total data).





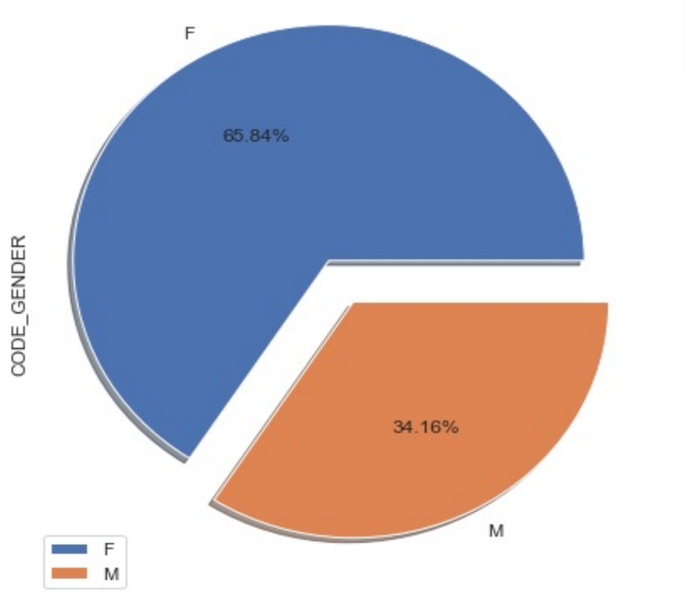
```
appdata.TARGET.value_counts(normalize=True)
```

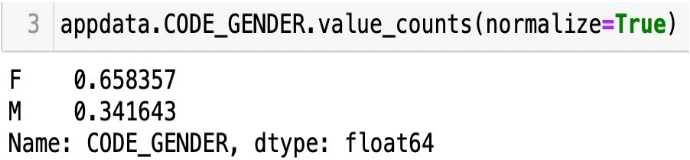
```
0    0.919271
1    0.080729
Name: TARGET, dtype: float64
```

Analysis

Imbalance Percentage

This Pie plot on gender indicates that out of the total data there are 65.84% females and 34.16% males under analysis in the current applications data.





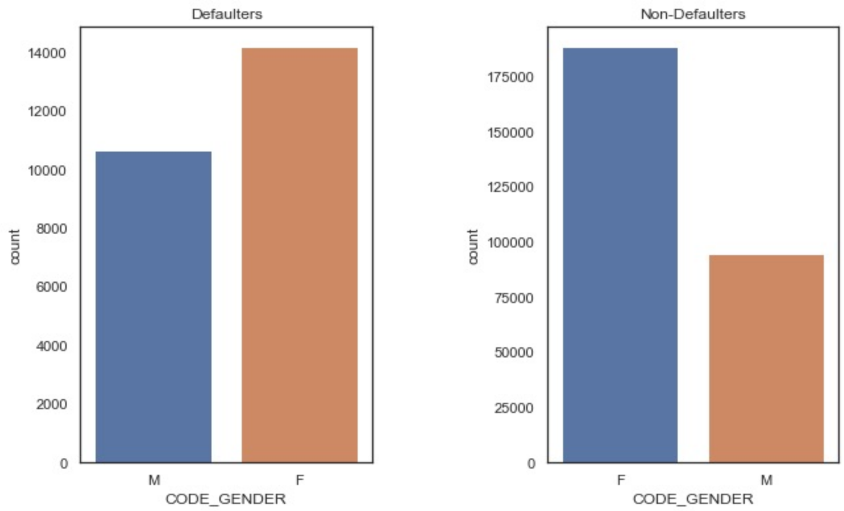
```
appdata.CODE_GENDER.value_counts(normalize=True)
```

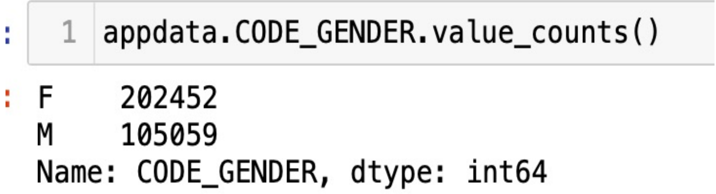
```
F    0.658357
M    0.341643
Name: CODE_GENDER, dtype: float64
```

Univariate Analysis

Target vs Gender

- The count for female defaulters are around 14000 whereas count for female non defaulters is around 185000.
- Male defaulters are around 11000 and male non defaulters count is around 98000 approx. So, the percentage of male defaulters are (11000/105059*_100=10.7%).Percentage of female defaulters is (14000/202452*_100=6.9%).
- The percentage of male non defaulters is approx.(100-10.7=89.3%) and female non defaulters=(100-6.9=93.1%)
- So overall data indicates the Males have more tendency to default a loan than females.

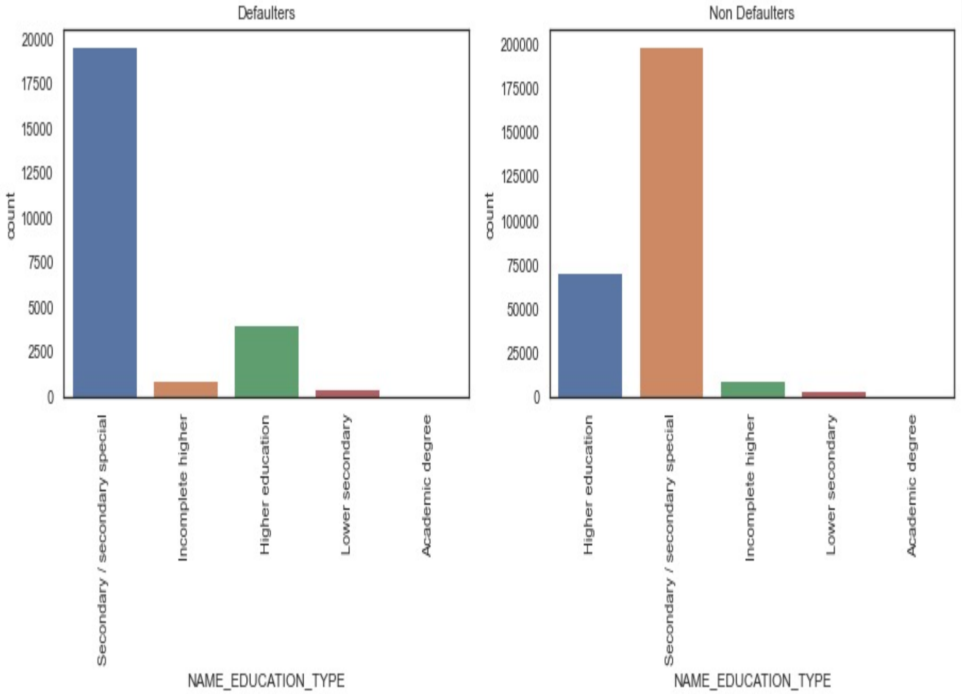




```
appdata.CODE_GENDER.value_counts()
```

```
F    202452
M    105059
Name: CODE_GENDER, dtype: int64
```

Univariate Analysis



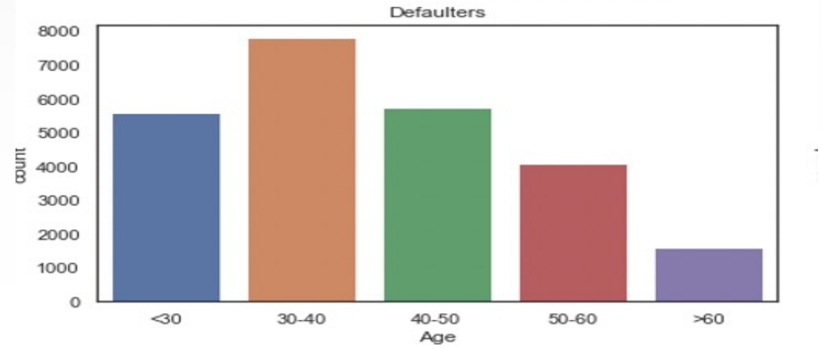
Target vs Education Type

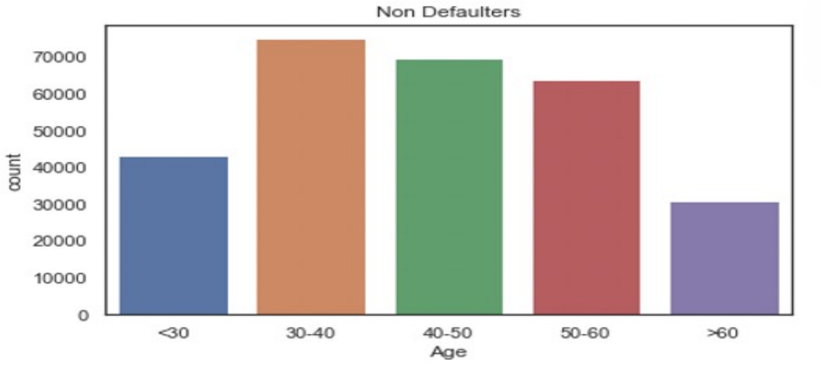
- It has been observed that the maximum defaulters are in education category of Secondary/Secondary Special which account to 79% of the total defaulters. This is the indication that people who have education till secondary level earn less or have low source of income and are not able to pay the loans on time
- Also the non defaulters are also from the education category of Secondary/Secondary Special which account to 70% of the total non defaulters.
- The analysis is very close for defaulters and non defaulters when it comes to Education Type

Univariate Analysis

Target vs Age Categories

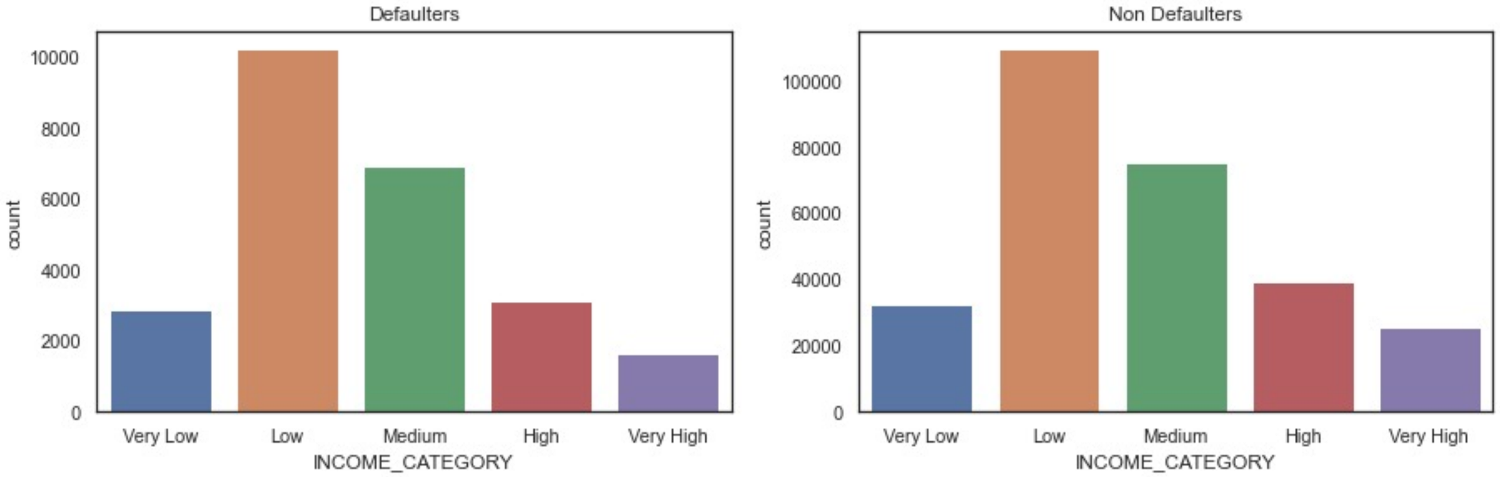
- It has been observed that nearly 31% of defaulters are in age group of 30-40 years followed by people in age group of 40-50 years. Also the graph indicates that these age group people are mostly defaulters.
- Also it has been observed that most non defaulters are in the age group of 30-40 years and 40-50 years.
- People of older age are very less in defaulters category, also the count for old age people in non defaulters is less. So we can say that old age people are the best option for the banks to provide loan.
- Even people of age less than 30 years are also less in defaulters and non defaulters.





Univariate Analysis

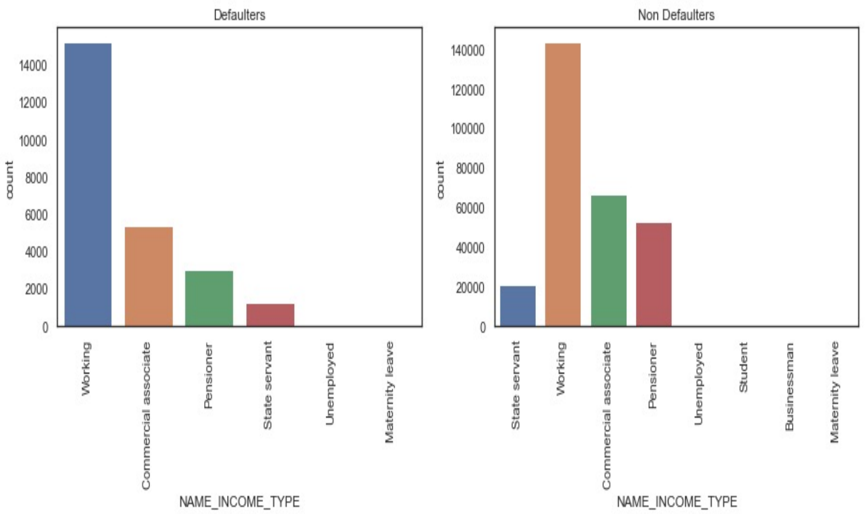
Target vs Income Category



- The plot indicates that the people with low income are the ones who are defaulters. This when compared with a real life scenario, the people with low salary are the ones in need of loan and mostly they are not able to payback the same. These low income group people are followed by medium, high, very low and very high income categories.
- On the other hand the non defaulters are also seen to be following the same trend as defaulters.
- In general it is all the payback capacity of a person that makes a defaulter or a non defaulter. A person who has low income is in need of loan, takes credit and may not be able to pay back on time, so that person becomes a defaulter. But at the same time a low earning person may need a loan in need and can also pay the loan on time as well, so he is a non defaulter.

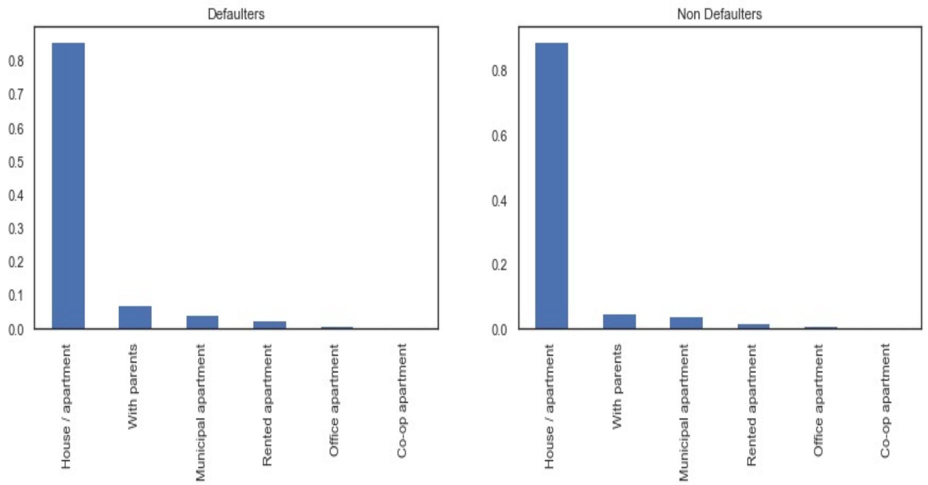
Univariate Analysis

Target vs Income Type



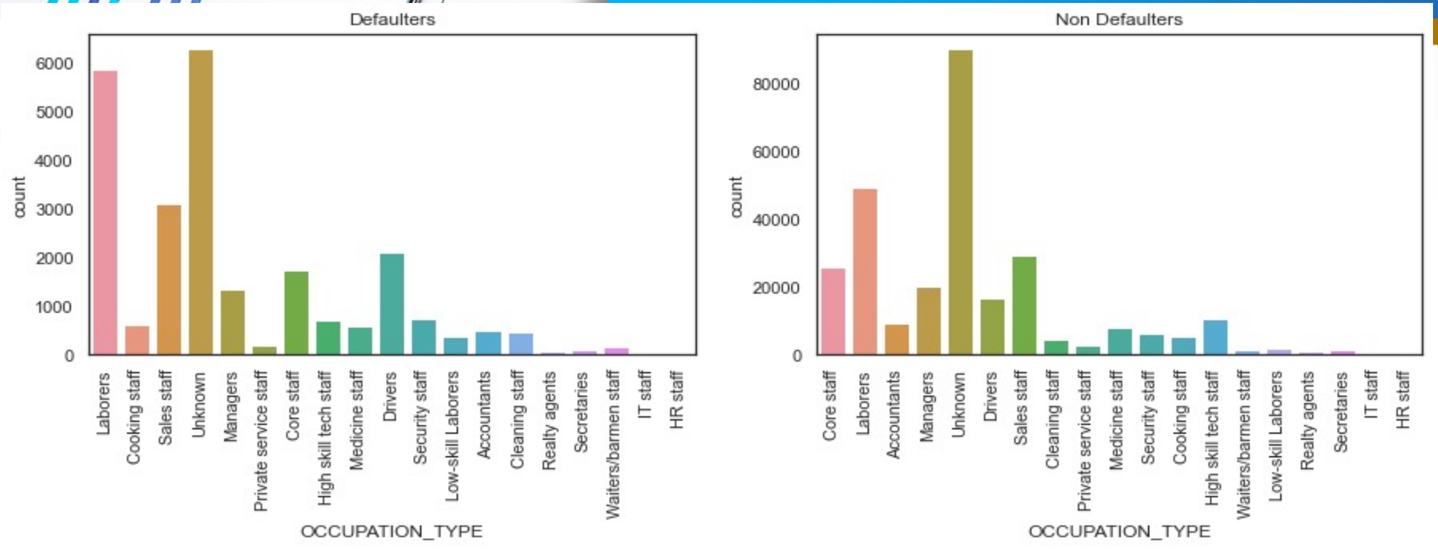
The plots indicate that maximum defaulters and non defaulters fall in category of working class people. Or in other words we can say, most loan applications are from working class people.

Target vs Housing Type



The data indicates that maximum people stay in their own house/apartments in both the defaulters and non defaulters category

Univariate Analysis



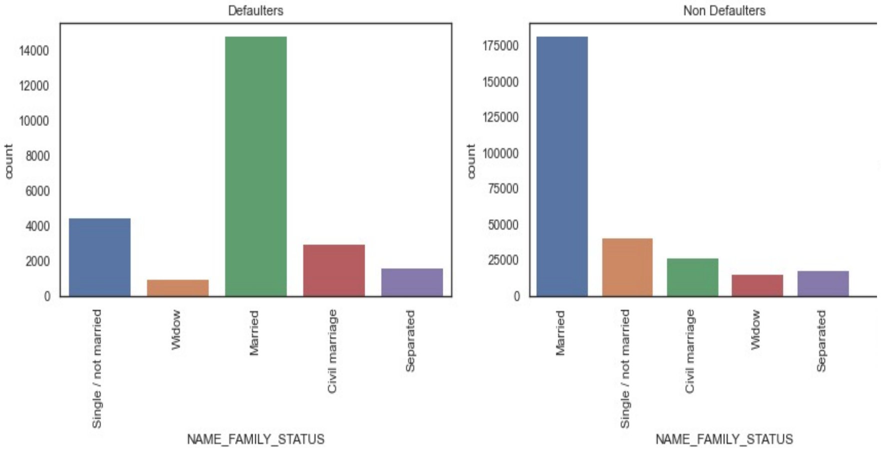
Univariate Analysis

Target vs Occupation Type

The plot indicates that the laborers occupation type is more in defaulters and non defaulters.

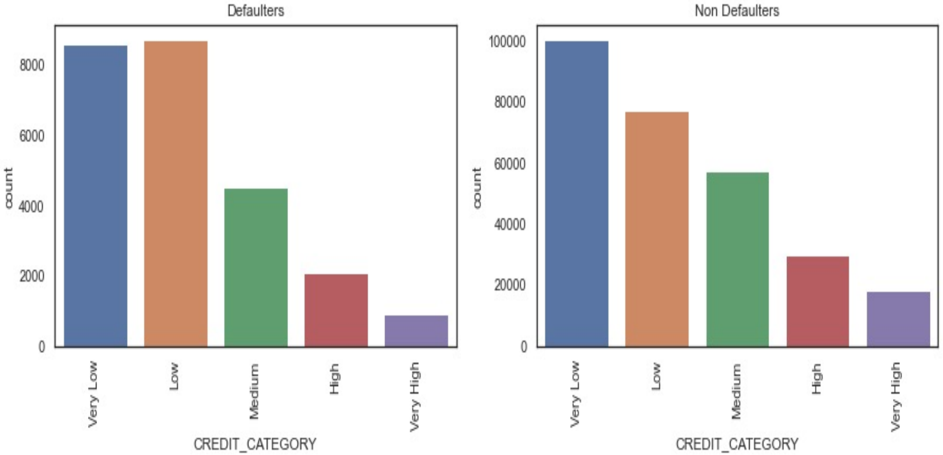
Univariate Analysis

Target vs Family Status



The graph indicates that most of the defaulters fall in married category. Also, the same is true for the non defaulters as well.

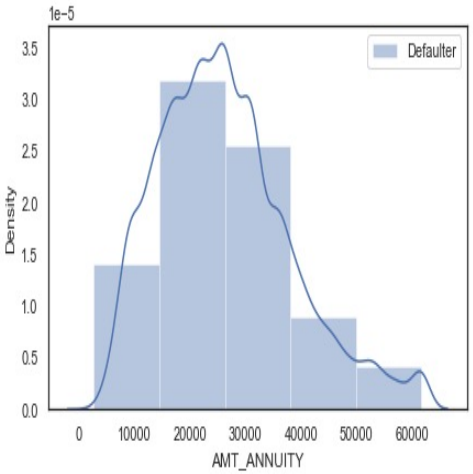
Target vs Credit Categories

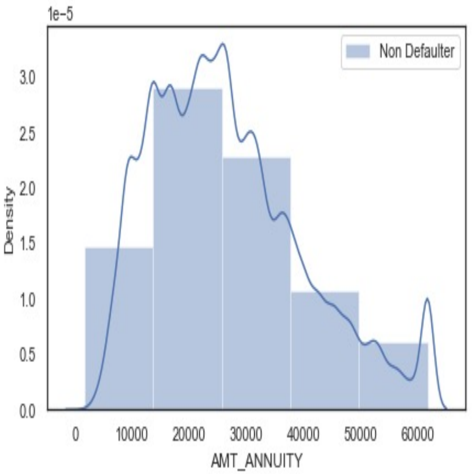


The graph clearly indicated that the people who have very low and low credit amount are more in number in defaulters. Also, the people who have very low credit are also high in count under non defaulters

Univariate Analysis

Annuity Analysis



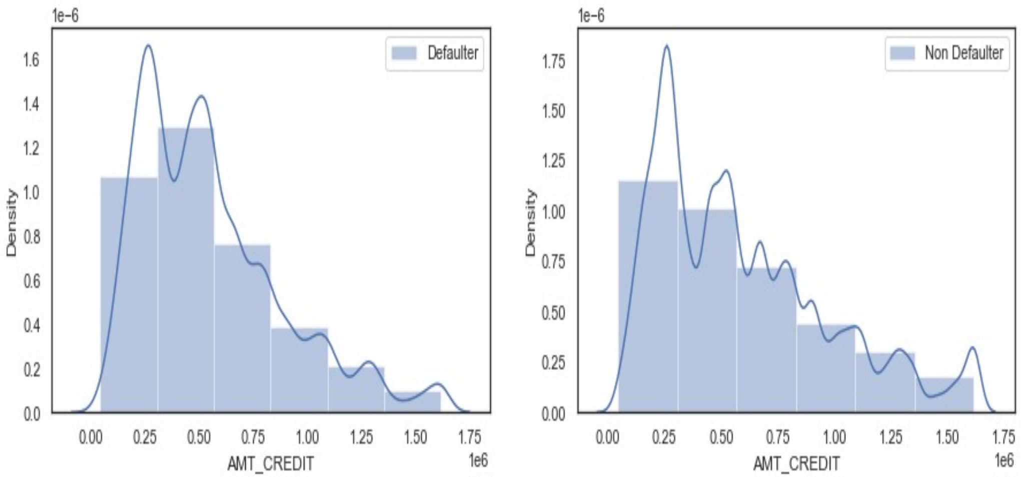


- The graph of annuity indicates that the spread of annuity is more between 15000 to 25000 and it gradually decreases over increase in annuity value. So, the people having annuity value between 15000 and 25000 are the ones who have defaulted the loan payment.
- On the other hand, in case of non defaulters there is a peak in density between annuity values of 15000 to 28000 and people in this annuity range are non defaulters.

Univariate Analysis

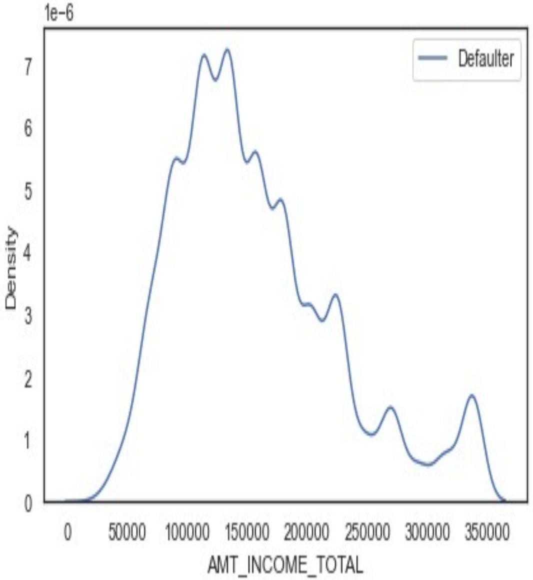
Credit Amount Analysis

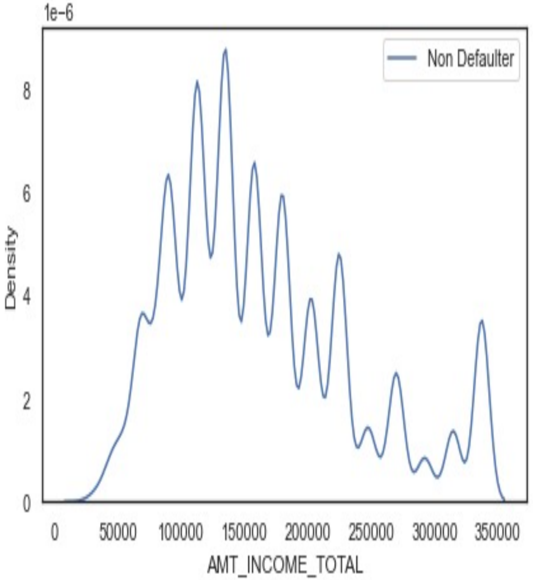
- Under defaulters the credit values have peaks at 250000 and 500000.
- In non defaulters the credit values have peaks at 250000 then at 500000 and followed by rise and falls and so on but the density keeps on decreasing post 500000.



Univariate Analysis

Income Analysis





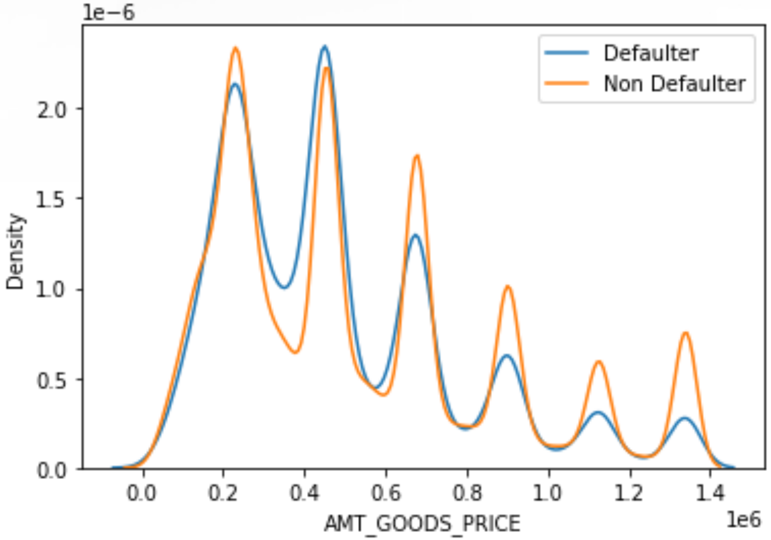
- It has been observed that maximum defaulters are those people in whose income is in range 1 to 1.5 lakhs. Also, as the income increases to 3 and 3.5 lakhs the count of defaulters decreases.
- On the contrary in non defaulters there is a mixed trend. People with high and low income group are all non defaulters..

Univariate Analysis

Goods Price Analysis

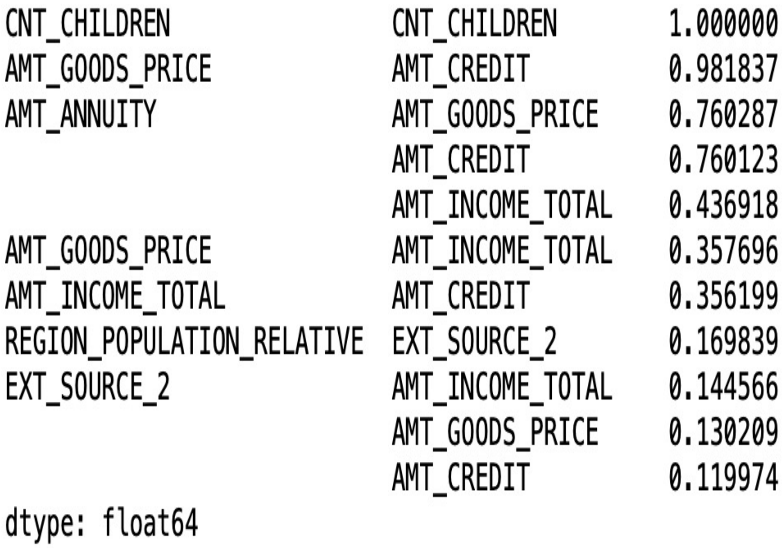
The curve indicates that both defaulters and non defaulters show a similar trend when compared on basis of price of goods.

This shows that the defaulters and non defaulters cannot be related to goods price



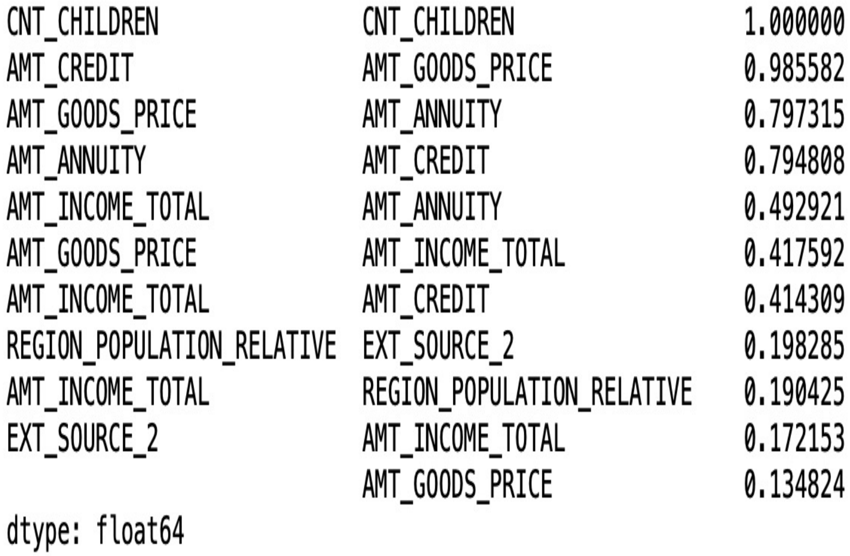
Bivariate Analysis

Correlation Matrix



```
CNT_CHILDREN               CNT_CHILDREN        1.000000
AMT_GOODS_PRICE            AMT_CREDIT          0.981837
AMT_ANNUITY                AMT_GOODS_PRICE     0.760287
                           AMT_CREDIT          0.760123
                           AMT_INCOME_TOTAL    0.436918
AMT_GOODS_PRICE            AMT_INCOME_TOTAL    0.357696
AMT_INCOME_TOTAL           AMT_CREDIT          0.356199
REGION_POPULATION_RELATIVE EXT_SOURCE_2        0.169839
EXT_SOURCE_2               AMT_INCOME_TOTAL    0.144566
                           AMT_GOODS_PRICE     0.130209
                           AMT_CREDIT          0.119974
dtype: float64
```

Defaulters

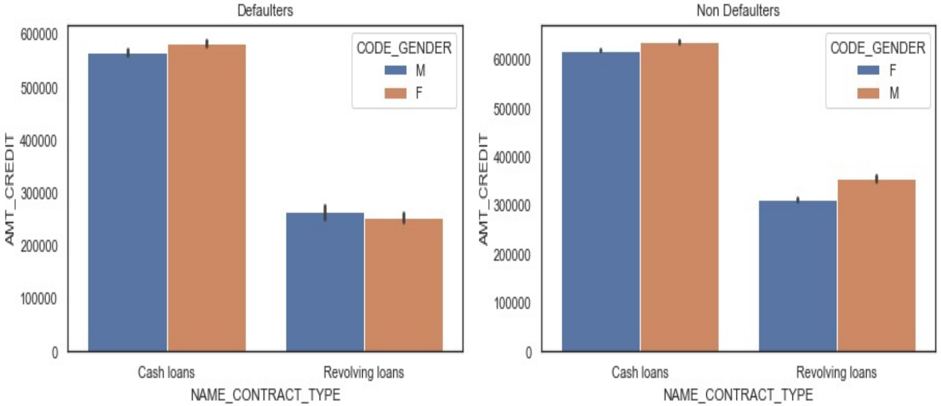


```
CNT_CHILDREN                CNT_CHILDREN                  1.000000
AMT_CREDIT                  AMT_GOODS_PRICE               0.985582
AMT_GOODS_PRICE             AMT_ANNUITY                   0.797315
AMT_ANNUITY                 AMT_CREDIT                    0.794808
AMT_INCOME_TOTAL            AMT_ANNUITY                   0.492921
AMT_GOODS_PRICE             AMT_INCOME_TOTAL              0.417592
AMT_INCOME_TOTAL            AMT_CREDIT                    0.414309
REGION_POPULATION_RELATIVE  EXT_SOURCE_2                  0.198285
AMT_INCOME_TOTAL            REGION_POPULATION_RELATIVE    0.190425
EXT_SOURCE_2                AMT_INCOME_TOTAL              0.172153
                            AMT_GOODS_PRICE               0.134824
dtype: float64
```

Non Defaulters

Bivariate Analysis

Gender vs Contract Type



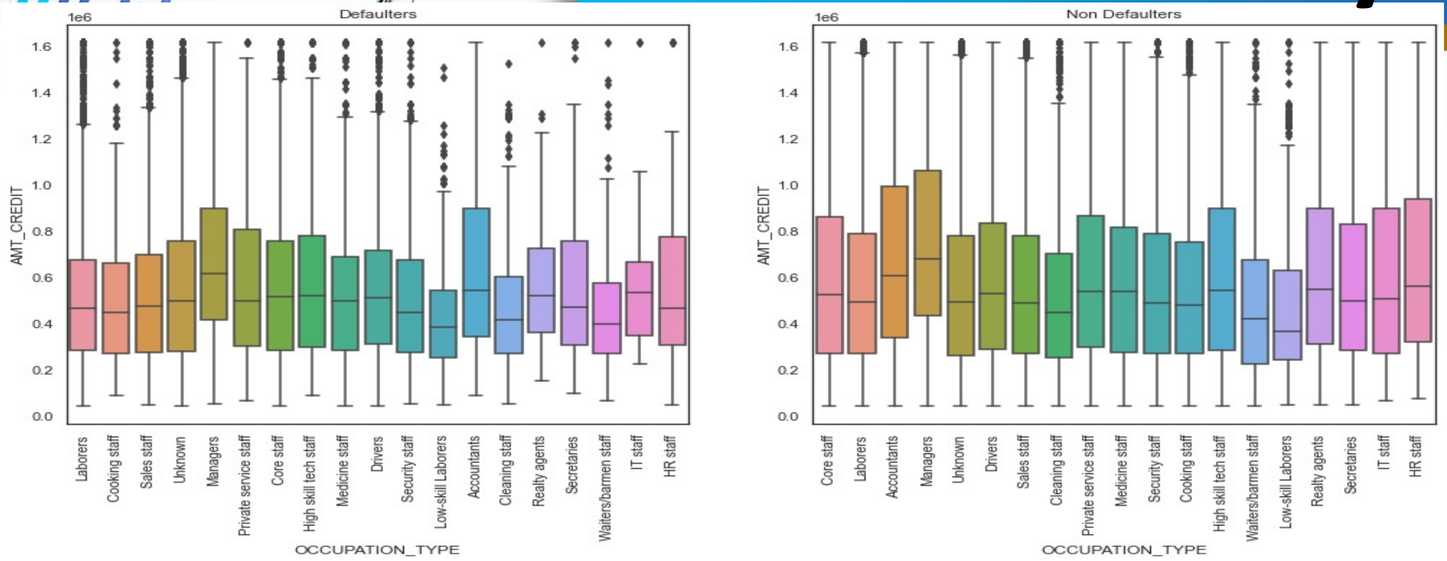
Insight: More cash loans are offered as compared to revolving loans

- The graph clearly indicates that there are more females who get cash loans as compared to males in the defaulters category. Whereas the revolving loans are offered more to men than females.
- In the non defaulters category the cash loans and revolving loans are more given to male and less to females.

So we can say that banks can trust males while giving cash loans and revolving loans as they are non defaulters. In case of defaulters banks should be more careful about females while giving cash loans and careful for males while giving revolving loans

Bivariate Analysis

Occupation vs Credit Amount



In defaulters

> it is seen that Accountants and Managers have come out to be ones who have more amount of credit with them and they are defaulters of loan too. So Banks and financial institutions should be more careful and alert in inspecting/interrogating the assets and paying capacity of managers and accountants while offering loans.
>
> In the same manner the next category of occupation that has more credit amount is Private service staff, high skill service staff followed by HR Staff, secretaries and drivers. As per analysis these categories tend to default the loan payment. So banks need to be careful for them
>
> The remaining categories are also defaulters but have less loan amounts with them

Non Defaulters

> It is evident that managers and accountants are the ones who have taken maximum loans and have paid the loans on time.

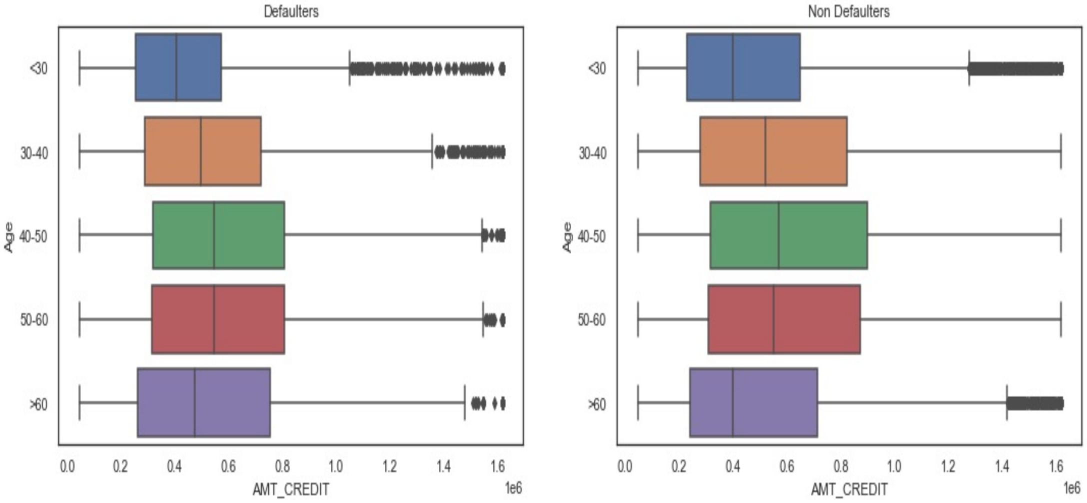
So in short the banks should be more vigilant for managers and accountants while giving loan.

Bivariate Analysis

- In defaulters maximum credit has been taken by people in age group 40-50 and 50-60 followed by people of age more than 60 years and then between 30-40 years
- People of age less than 30 years have low credit amount with them.
- In non defaulters category it is evident that the middle aged people i.e. in age group of 40-50 years tend to take more loans followed by old people who are 50 and above then followed by young adults in age of 30-40 years.

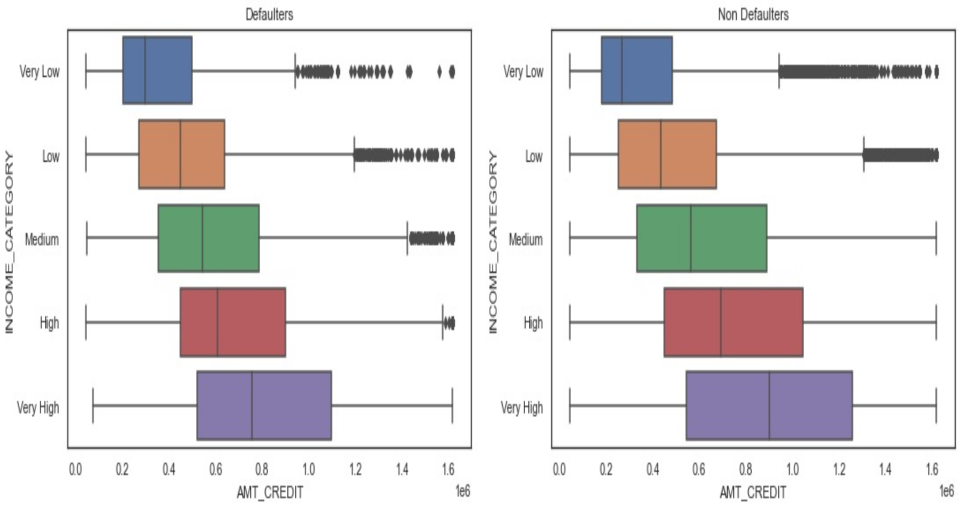
From the non defaulters data it can be recommended that people who are less than 30 years take less credit amount and also return the amount as well, so they are trustworthy, also the people of age more than 60 an be trusted by banks in providing loans as they are non defaulters as well. Generally old people have good amount of savings to payback the loans.

Age vs Credit Amount



Bivariate Analysis

Income Category vs Credit Amount

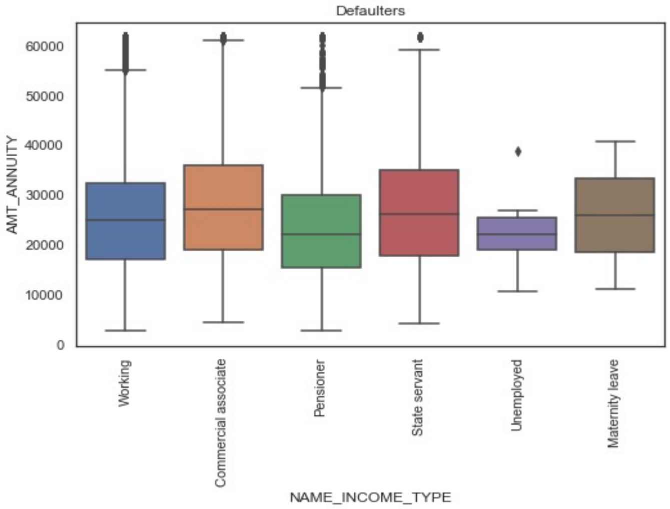


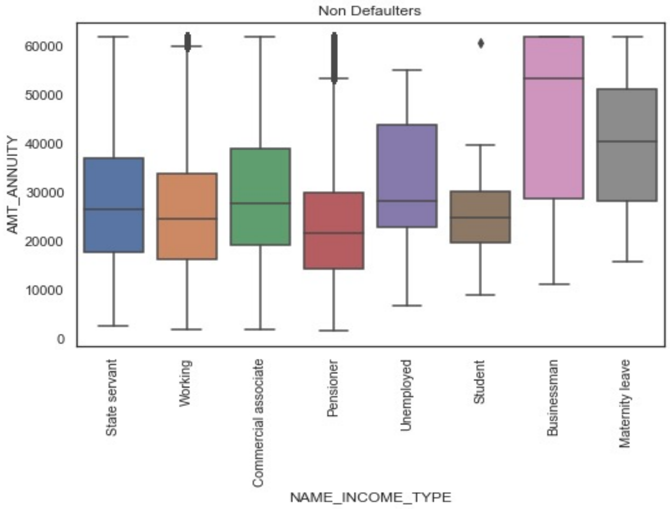
- People who are defaulters with high very high income have more credit amount and the people with low income have low credit. This also indicated that the way the income of defaulters increase so is the increase in the credit amount provided to them. This is quite evident from the median in the defaulter's graph.
- On the other hand, the same trend is seen in non defaulters as well. But the difference is that the non defaulters with high and very high income have been given more credit (amount as high as 13 lakhs) as compared to the people in high and very high-income group who fall in defaulters category get around 11 lakhs)

From above it is evident that bank has developed trust on certain people as they are non defaulters and are awarded more credit. Whereas the same trust is not for the defaulters in the same income category.

Bivariate Analysis

Annuity vs Income Type





The people who are defaulters, it has been observed that

- Commercial associates and state servants that have a similar Annuity amount. The same is the case for the people who are on maternity leave as well. These 3 categories: commercial associates, State servants and those on maternity leave have high annuity assets of value around 35000, a median of 25000 and lower quartile value at around 20000.
- Also, the unemployed people are the ones who have least annuity amount though the median value is somewhere between 20000 to 25000.
- Pensioners being the old people in the income category definitely have more annuity as compared to unemployed people and they share the same median as the unemployed people in the defaulters category.
- The working people fall in the middle bracket . They have annuity a little above 30000 and median value slightly above the median for pensioners and unemployed. So, these are better off category of people.

The defaulters analysis of Annuity and Income type show us that the Commercial associates, State Servants and people on Maternity leave have high annuity amounts, followed by the working class, pensioners and unemployed respectively

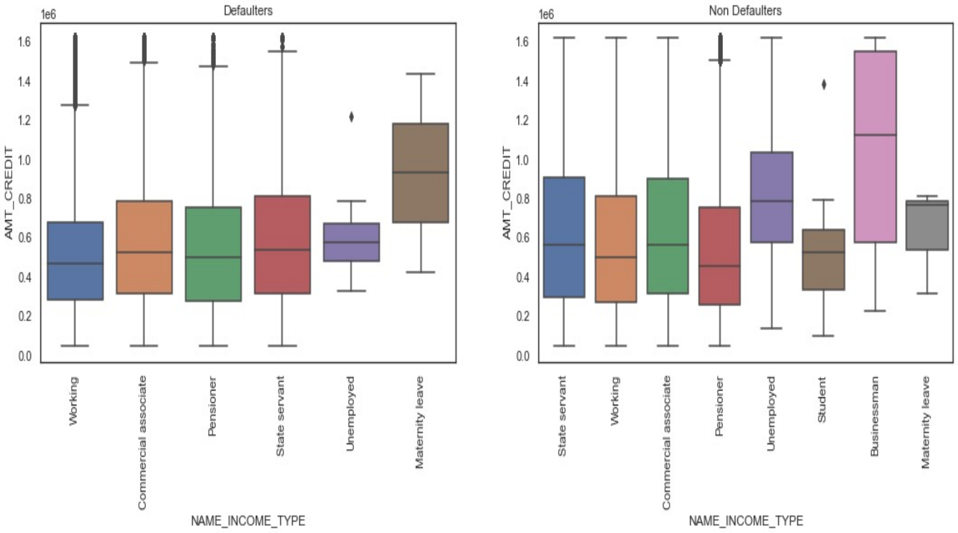
People who are non defaulters, it has been observed that

- Firstly, the businessman are the ones who lead in having high Annuity amount and are non defaulters as well as per the data.
- The median value for commercial servants and unemployed non defaulters is at the same value of approx 28500. But interestingly the unemployed who are non defaulters have more annuity with them of amount around 45000. This means these unemployed non defaulters have good savings and good investments as compared to the commercial servants
- On the same lines as above point, it is evident that the working class people and students also have same median annuity value, but working people have more annuity amount as compared to students. This can be for obvious reasons that students have only source of income as pocket money or could be internship which is way less than the income of a working class person. So the non defaulter working people have more annuity value.
- Non defaulter State servants also enjoy a good annuity amount followed by the least in the category which are the pensioners
- Interesting to note that maternity leave non defaulters have high annuity amounts.

From the above analysis we can say that the annuity and income type relation between defaulters and non defaulters independently has its own variations. eg Observing the defaulters and non defaulters under each income type they have similar annuity amounts.

Bivariate Analysis

Credit Amount vs Income Type



Defaulters:

- Under defaulters it has been seen that people take more loans i.e. have more credit amount who are on maternity leave
- On the other hand commercial associates and state servants have credit values.
- The working people have low credit amount under defaulters category
- Pensioners also have loans amounts in the same category as commercial associates and state servants

The main reasons behind these can be interpreted as, people who plan a children and go on maternity leave, take more credit as the process involves a lot of finance.

Also, the amount of loan taken by Commercial associates, state servants or working-class people are of lesser amount which is between 3lakhs to 7 lakhs approximately.

Non defaulters:

- Under non defaulters the businessman are the ones who have high credit amounts in their kitty
- Business people are followed by unemployed people, who need loans for their living, but they are also non defaulters. They pay their loan.
- Just like defaulters state servants and commercial associates have same credit pattern followed by the working people
- However, unlike pensioners in defaulters, the non defaulter pensioners take less credit.

The businessmen require more credit to run their business, so they have high loan amounts. The unemployed people need loan to meet their ends, so they take loans. Pensioners take less credit as they have their own savings already, but to meet certain uncertainties they make take some loan, so their credit amount is less.

Bivariate Analysis

Defaulters:

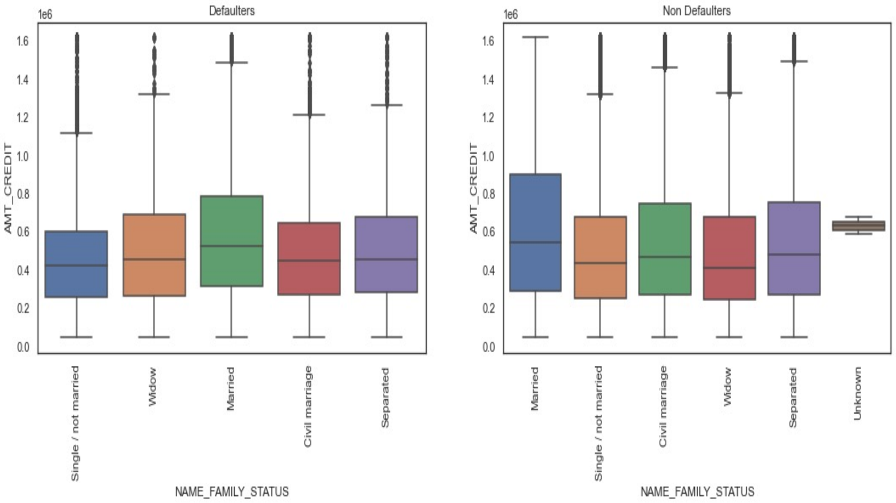
- The boxplot indicates that married people are the ones who take more credit and are defaulters as well
- The median value for the defaulter widows and the separated category have same loan amounts in their kitty
- People who are single or have done civil marriage in the defaulters category are the ones who have low credit amounts ranging between 3 lakhs to 6 lakhs.

Non-Defaulters:

- In Non defaulters it has been seen that married people have maximum credit amount with them
- Those who have done civil marriage or are separated are the ones who have same credit amount taken and they have median values also same.
- Similarly single and widow category in family status have same credit amounts taken ranging between 3 lakhs to a little higher than 6 lakhs.

Looking at the plots it is evident that there are less widows as defaulters and also some widows as non defaulters. So the first option for banks to provide loans is to widows

Credit Amount vs Family Status

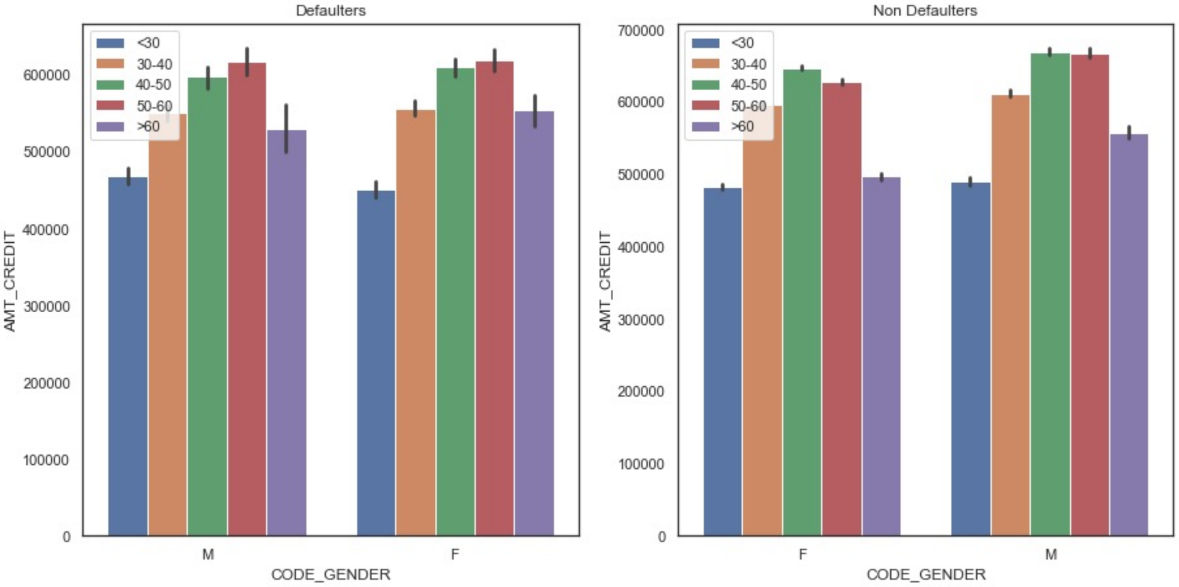


Bivariate Analysis

- In defaulters category it is quite evident that male aged between 50-60 and 40-50 take more credit amounts and are defaulters as well. In females also, age groups of 50-60 and 40-50 take more credit followed by the old aged (>60) and young adult (30-40) category.
- In non defaulters maximum loan is with middle aged (40-50) male and female followed by old aged and young adults.

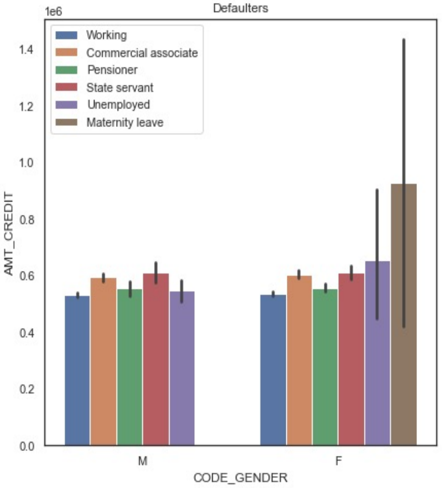
The most safe bet for bank to provide loans can be Young adult male and females as there are less defaulters in this category but with thorough checks and background scrutiny

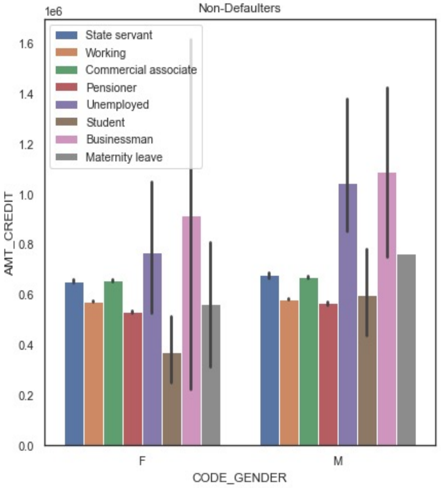
Gender vs Age Category vs Credit



Bivariate Analysis

Gender vs Age Category vs Credit





- Defaulters:
  - The females on maternity leave tend to take higher credit and are the major defaulters as well.
  - Whereas defaulter males it is seen that the ones whose income source is sate or are working as state servants and those who are commercial associates take loan of high amount.

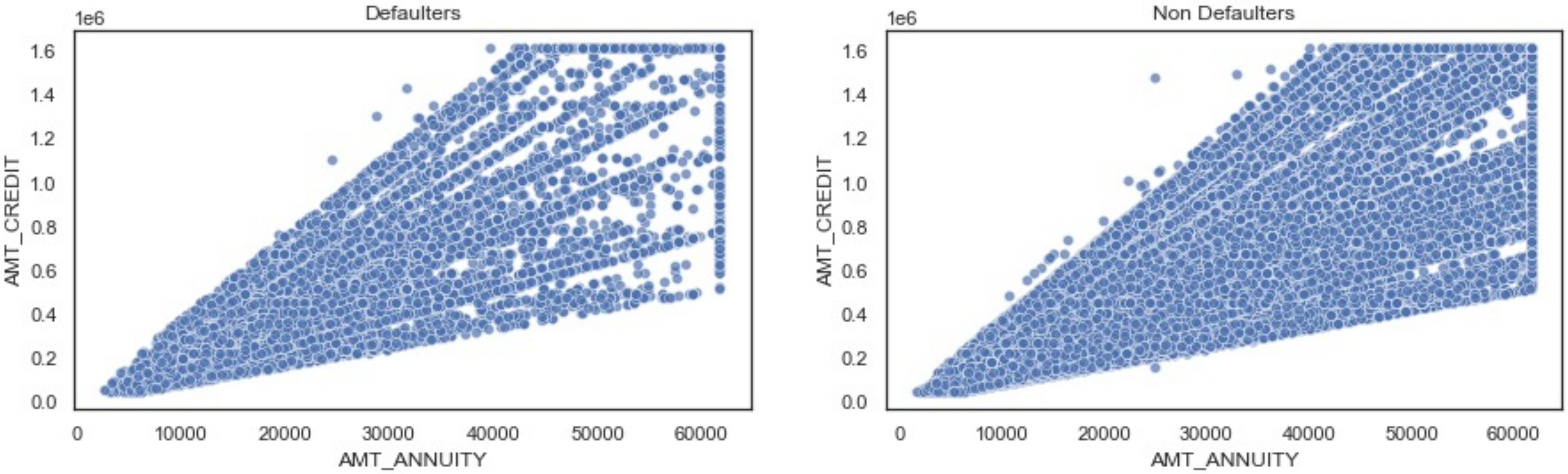
So, in defaulters category, banks should be careful while giving loan to pregnant women.

- Non Defaulters:
  - From the plot it is evident that the businessmen whether male or female are the ones who take more loans and are also non defaulters. So, these are the safe bet for banks to provide loan to.

Also an important thing to note is that businessman and student category are not in defaulters category.

Bivariate Analysis

Annuity Amt vs Credit Amount

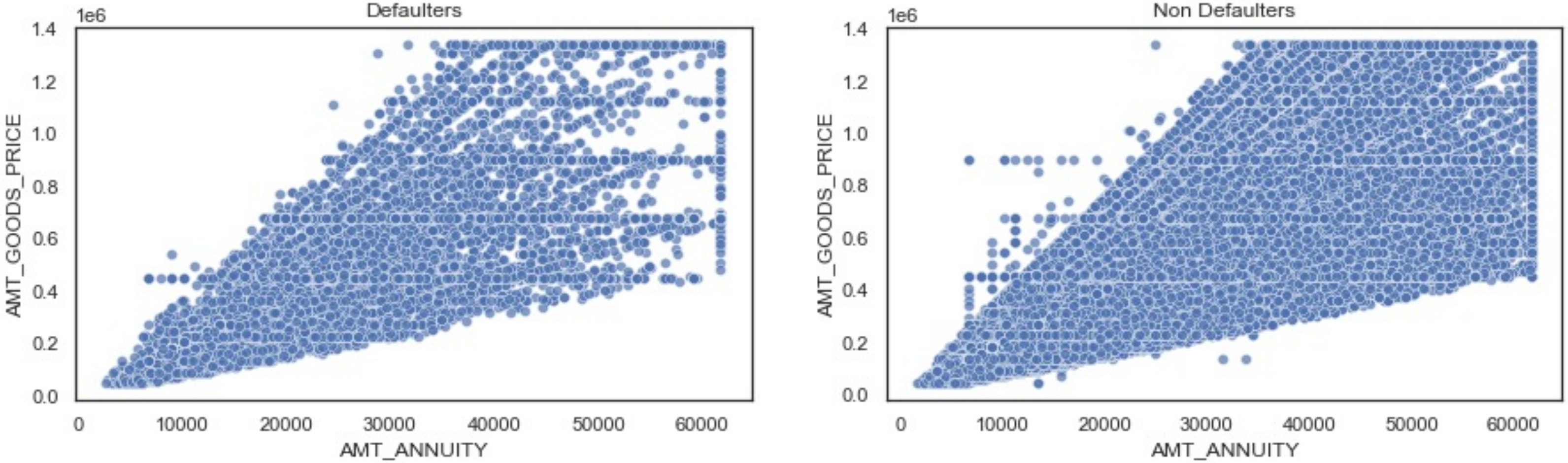


```
Correlation b/w Annuity and Credit for defaulters 0.7601234136216823
Correlation b/w Annuity and Credit for non defaulters 0.7948078743243163
```

The scatterplots between Annuity and Credit indicate that they have positive correlation and have linear correlation between the two variables. A person with more Annuity assets has better chance to get higher credits and vice versa

Bivariate Analysis

Annuity Amt vs Goods Price

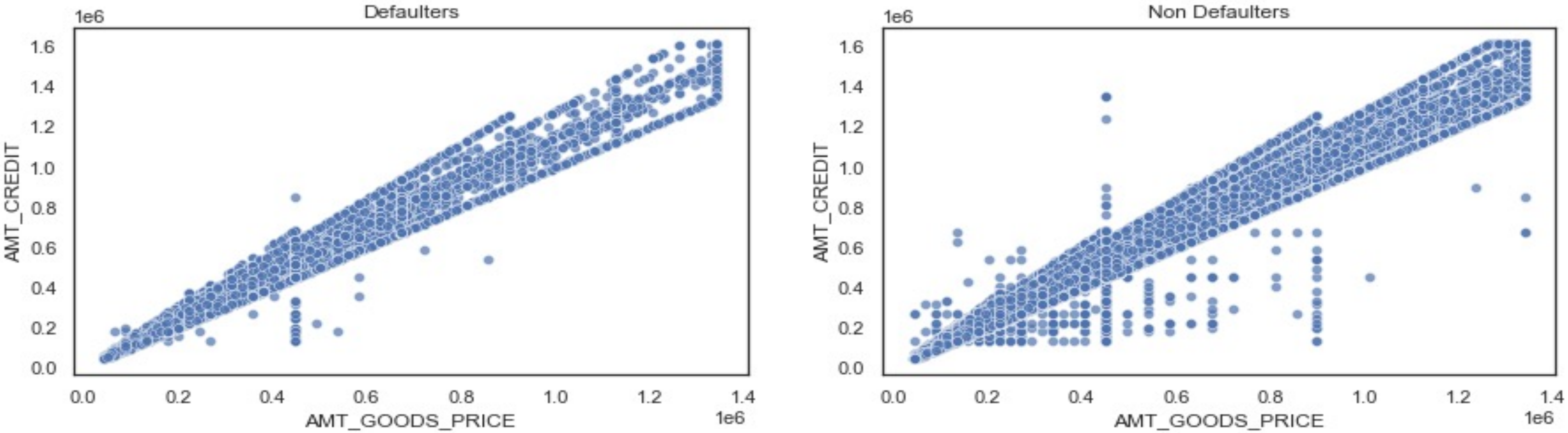


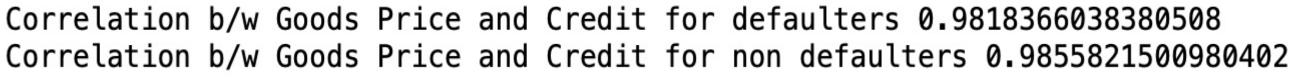
```
Correlation b/w Annuity and Goods Price for defaulters 0.7602866472620408
Correlation b/w Annuity and Goods Price for non defaulters 0.7973154338305387
```

The scatterplots between Annuity and Good Price indicate that they have positive correlation and also some linear correlation is also seen between the two variables. But one is not the cause for other. That is there is no causation. Increase in goods price does not increase the Annuity, even though there is strong correlation between them.

Bivariate Analysis

Goods Price vs Credit Amount



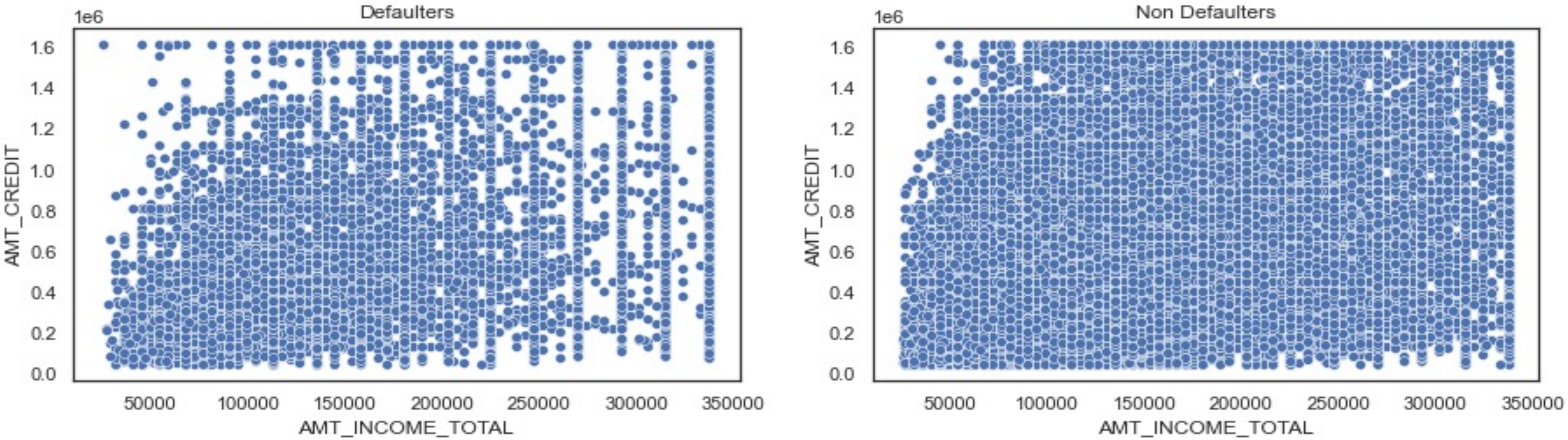


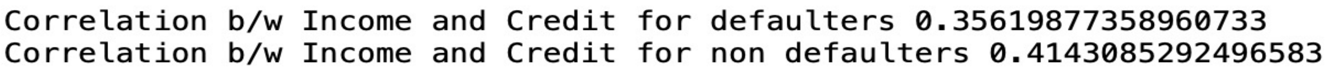
```
Correlation b/w Goods Price and Credit for defaulters 0.9818366038380508
Correlation b/w Goods Price and Credit for non defaulters 0.9855821500980402
```

There is very high correlation between Goods Price and Credit for defaulters. In Non defaulters also the correlation is positive and linear, but some outliers are also there. The correlation of 98% is quite high for the two variables. So overall we can say that as the goods price increase the credit amount also increases.

Bivariate Analysis

Income vs Credit Amount





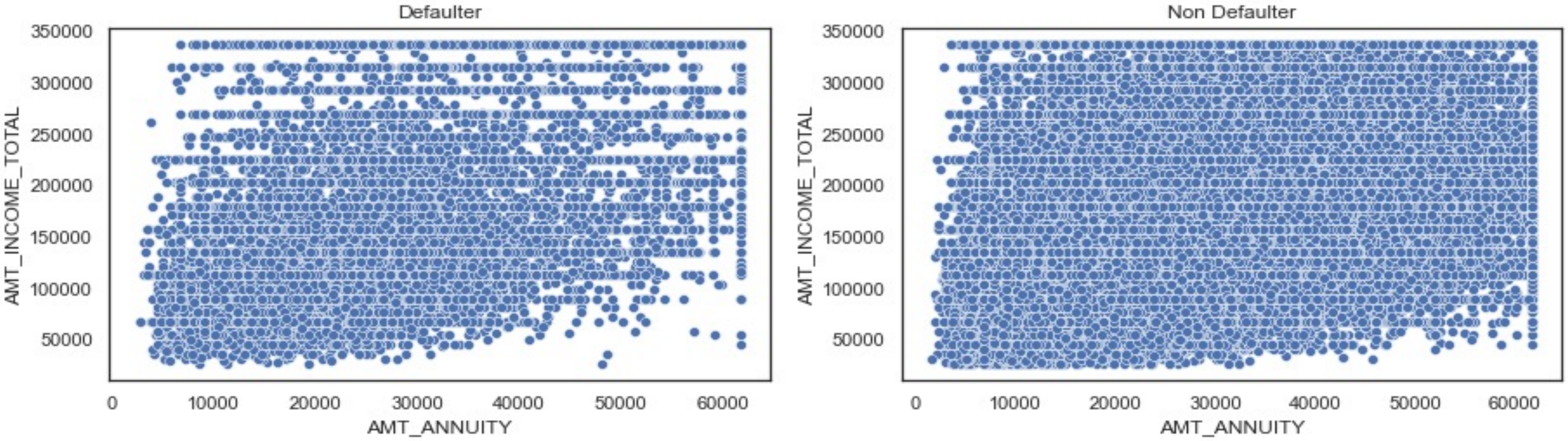
```
Correlation b/w Income and Credit for defaulters 0.3561987735896073
Correlation b/w Income and Credit for non defaulters 0.4143085292496583
```

The Defaulters data indicate that, the corelation between the income and credit is not that great. It is a positive correlation but here is no linear relation between them

The non defaulters scatter plot also indicates that there is no linear relation between Income and Credit

Bivariate Analysis

Income vs Annuity Amount



```
Correlation b/w Income and Annuity for defaulters 0.43691820152004285
Correlation b/w Income and Annuity for non defaulters 0.4929207394232968
```

The Defauters data indicate that, the corelation between the income and annuity is not that great. It is a positive correlation but here is no linear relation between them

The non defaulters scatter plot also indicates that there is no linear relation between Income and Anuuity

Multivariate Analysis

Housing Type vs Family Status vs Credit Amount

Defaulters:

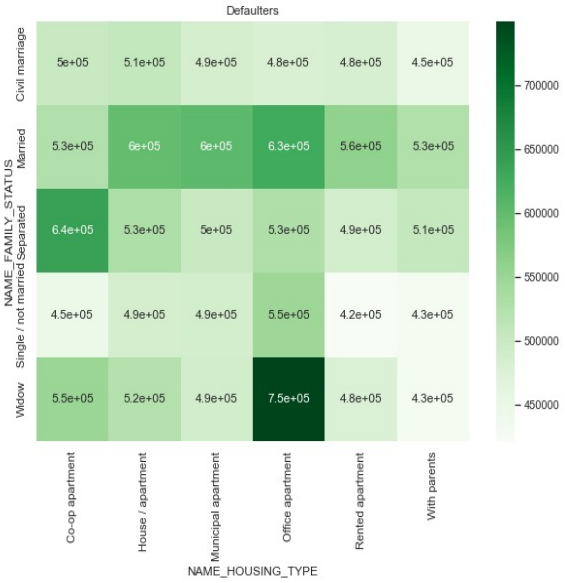
The heatmap for defaulters indicate that widows and office apartment housing type has a vey strong correlation, followed by separated and Co-op apartment.

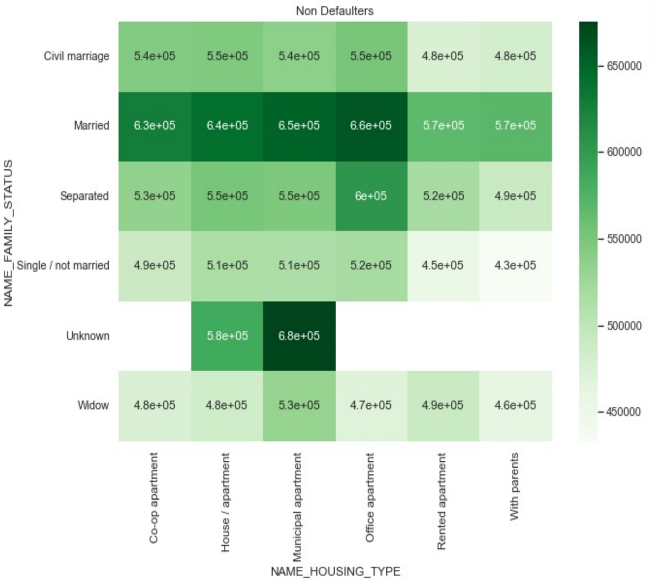
One more thing to be observed is that married people have a positive correlation with all the housing types and also a strong relation we can say by the colour of the boxes.

Non defaulters:

The strongest correlation exists between married people and Office Apartment.

In fact married people have strong correlation with all the housing types





Multivariate Analysis

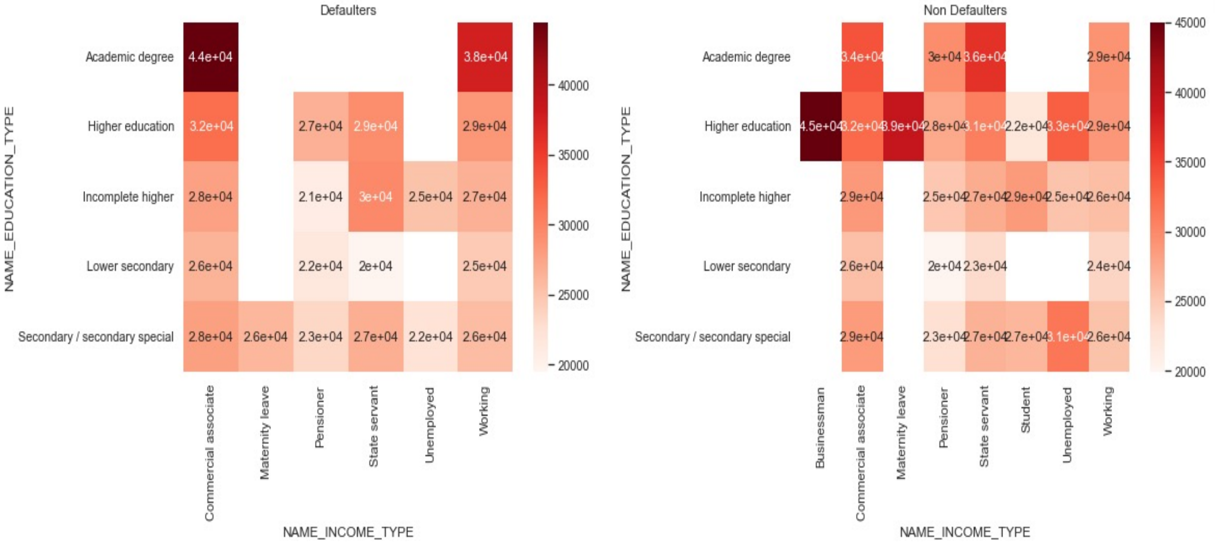
Education Type vs Income Type vs Annuity Amount

Defaulters:

The heatmap shows that the relation between academic degree and commercial associate is highly correlated whereas the lower sec is least correlated with any of the Income types

Non Defaulters:

We can depict from the graph that the higher education is the favourably corelated with Businessman



Previous Data Analysis

The analysis of the previous applicants has been done by merging the data frames. The data frames have been merged as follows:

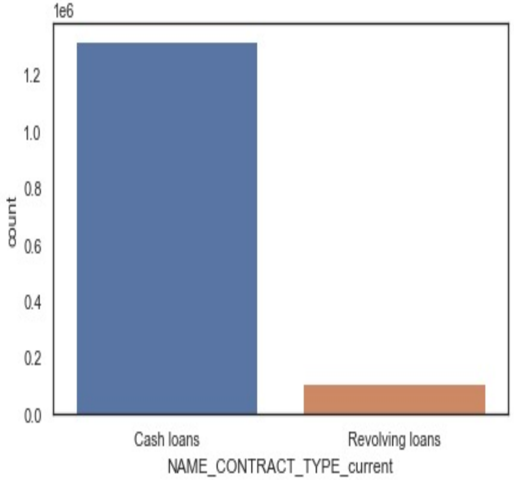
- Previous application data merged with the total data of application data frame
- Previous application data merged with only defaulters dataframe subset.

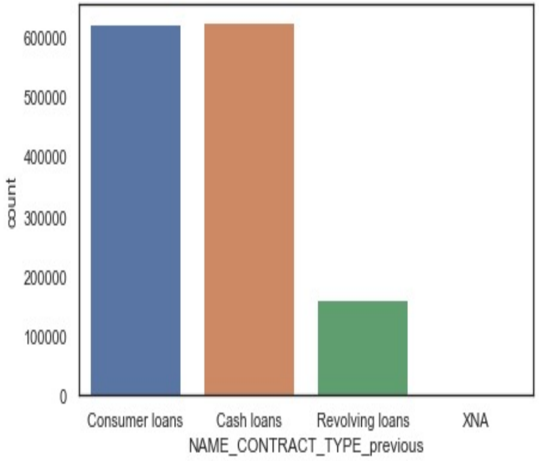
Previous Data Analysis

Analysis of Previous application dataframe (after cleaning) with the current application dataframe

- The plots indicate that the previous data base has three type of loans: Cash Loans, Revolving Loans and Consumer Loans. But in current application data set, there are only cash loans and revolving loans.

The data also indicates that in current applications there are more cash loans that are being provided to the people and in previous data there are more consumer loans that are provided to the people

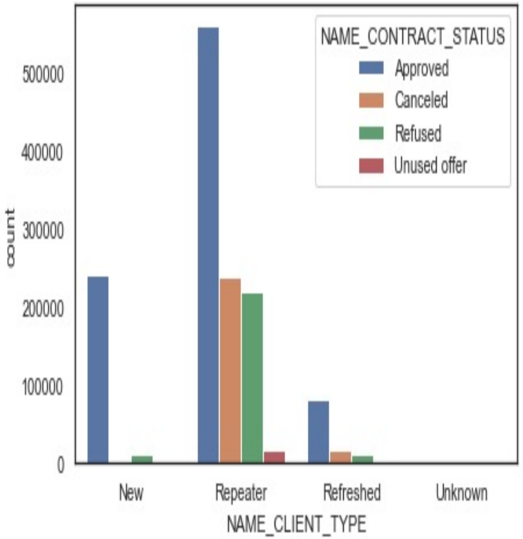


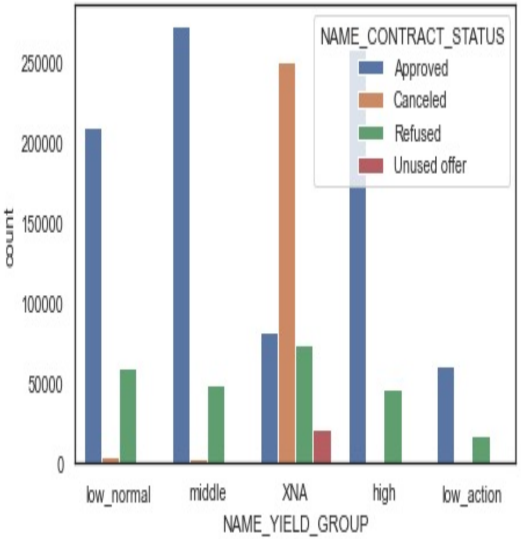


Previous Data Analysis

Analysis of Previous application dataframe (after cleaning) with the current application dataframe

- The above plots indicate that the people who have repeatedly applied for loan, their loans have been approved and the count of such approved loans is maximum.
- The people whose yield group is middle are the ones whose loans have been approved the most followed by the ones in high yield group.



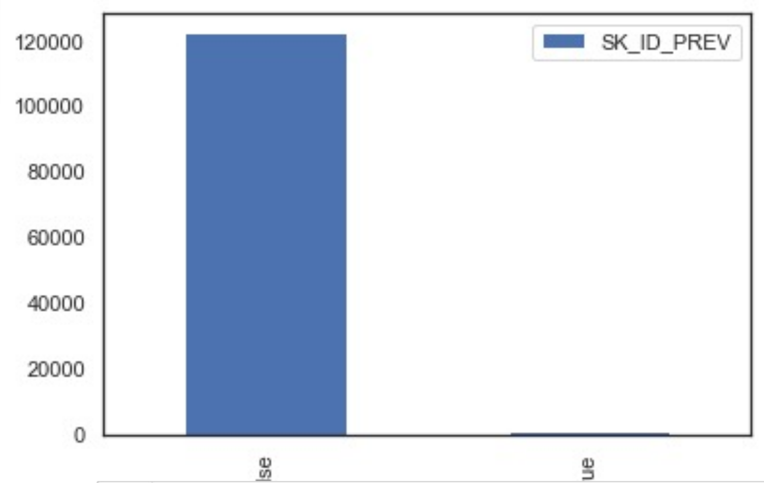


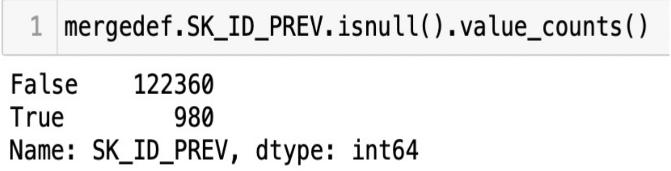
Previous Data Analysis

Analysis of Previous application dataframe (after cleaning) with the current application defaulters(Target=1) sub dataframe

- The values above indicate that in the previous data we have 980 missing values of previous id corresponding to the defaulters data. So from previous data comparison there is an addition of 980 people in defaulters category presently.

- Also there are 122360 applicants in the previous data who are still defaulters as per the current data application set. So 122360 people who were in previous application data set are still facing difficulties in making timely payments



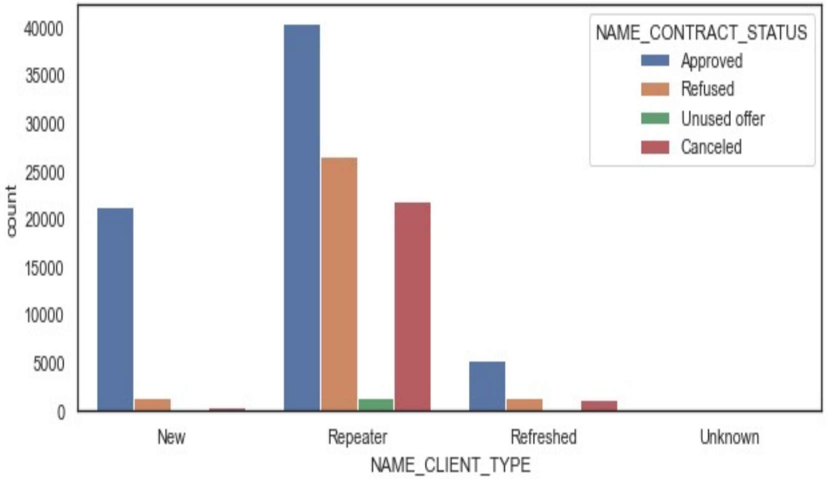


```
mergedef.SK_ID_PREV.isnull().value_counts()
```

```
False    122360
True        980
Name: SK_ID_PREV, dtype: int64
```

Previous Data Analysis

Analysis of Previous application dataframe (after cleaning) with the current application defaulters(Target=1) sub dataframe



The observation from the above countplot is that there are more repeaters from the previous data that are repeated in the current data. So the ones who had applied for loan previously and are defaulters at present.

Summary

- The analysis of both the data sets indicate that, in current applications cash loans are provided more and previously consumer loans were more.
- In the non defaulters there are businessmen and students which are not there in defaulters, so this indicates that banks trust businessmen and students in giving loans.
- In a general trend it has also been identified that men are more liable to default a loan than females.
- Also the people who have more income get more credit and also people with more income are less in defaulters.
- Pensioners are again one of the option that banks can rely on in giving loans.
- Women on maternity leave have found to be the top most defaulters. So banks should be careful in this category.
- Goods price and credit show a very high and positive correlation.
- Annuity and goods price are positively correlated but there is no causation between the two.

THANK YOU